Implication of AI driven classification models on metabolomics profiling dataset: Breast cancer case study

Journey through Data Exploration, Preprocessing, and Modeling

Presented by:
Dr. Ashish Sharma ,
Pr. Project Scientist,
Translational Bioinformatics Group, ICGEB

Remember the number

_ _ _

One

Two

Zero

Eight

One

Two

Eight

Three

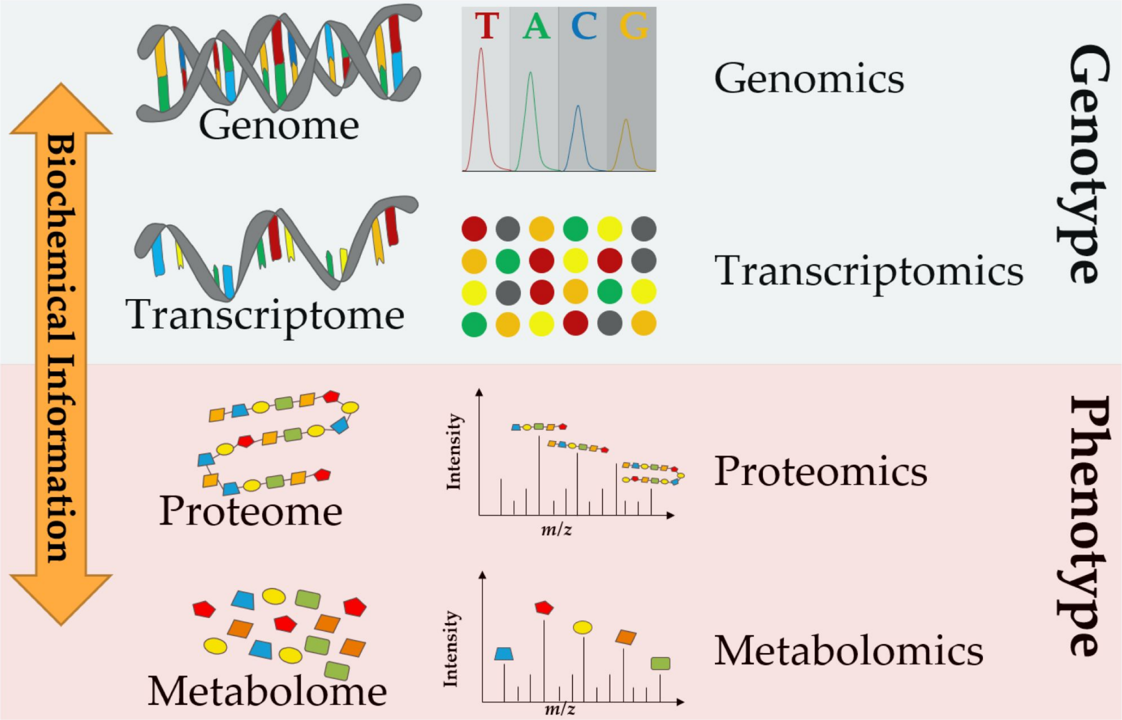
Zero

One

One

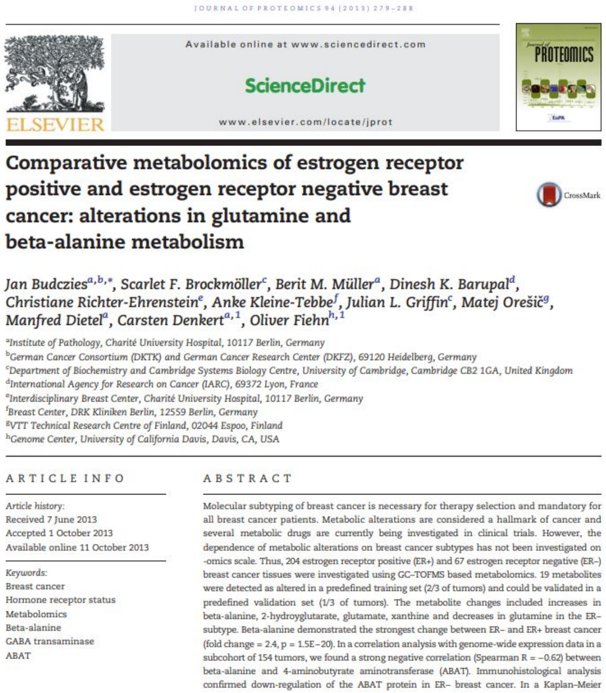
Introduction

Metabolomics is a field that explores small molecules in biological systems.



Comparative metabolomics of estrogen receptor positive and estrogen receptor negative breast cancer: alterations in glutamine and beta-alanine metabolism

Deep Learning Accurately Predicts Estrogen Receptor Status in Breast Cancer Metabolomics Data



JOURNAL OF PROTEOMICS 94 (2013) 279–288

Available online at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/jprot

ELSEVIER

Comparative metabolomics of estrogen receptor positive and estrogen receptor negative breast cancer: alterations in glutamine and beta-alanine metabolism

Jan Budczies[a,b,*], *Scarlet F. Brockmöller*[c], *Berit M. Müller*[a], *Dinesh K. Barupal*[d], *Christiane Richter-Ehrenstein*[e], *Anke Kleine-Tebbe*[f], *Julian L. Griffin*[c], *Matej Orešič*[g], *Manfred Dietel*[a], *Carsten Denkert*[a,1], *Oliver Fiehn*[h,1]

[a]*Institute of Pathology, Charité University Hospital, 10117 Berlin, Germany*
[b]*German Cancer Consortium (DKTK) and German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany*
[c]*Department of Biochemistry and Cambridge Systems Biology Centre, University of Cambridge, Cambridge CB2 1GA, United Kingdom*
[d]*International Agency for Research on Cancer (IARC), 69372 Lyon, France*
[e]*Interdisciplinary Breast Center, Charité University Hospital, 10117 Berlin, Germany*
[f]*Breast Center, DRK Kliniken Berlin, 12559 Berlin, Germany*
[g]*VTT Technical Research Centre of Finland, 02044 Espoo, Finland*
[h]*Genome Center, University of California Davis, Davis, CA, USA*

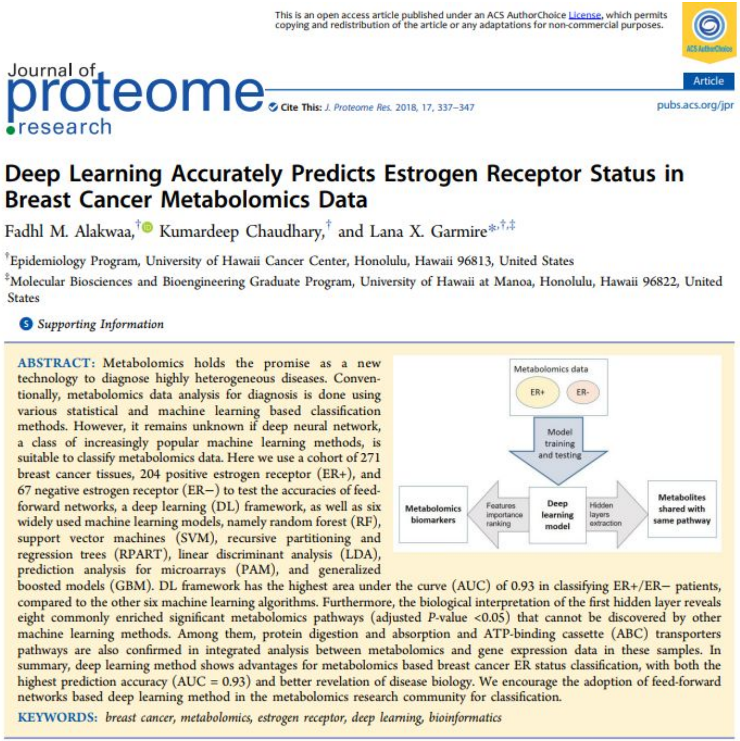
ARTICLE INFO

Article history:
Received 7 June 2013
Accepted 1 October 2013
Available online 11 October 2013

Keywords:
Breast cancer
Hormone receptor status
Metabolomics
Beta-alanine
GABA transaminase
ABAT

ABSTRACT

Molecular subtyping of breast cancer is necessary for therapy selection and mandatory for all breast cancer patients. Metabolic alterations are considered a hallmark of cancer and several metabolic drugs are currently being investigated in clinical trials. However, the dependence of metabolic alterations on breast cancer subtypes has not been investigated on -omics scale. Thus, 204 estrogen receptor positive (ER+) and 67 estrogen receptor negative (ER–) breast cancer tissues were investigated using GC–TOFMS based metabolomics. 19 metabolites were detected as altered in a predefined training set (2/3 of tumors) and could be validated in a predefined validation set (1/3 of tumors). The metabolite changes included increases in beta-alanine, 2-hydroyglutarate, glutamate, xanthine and decreases in glutamine in the ER– subtype. Beta-alanine demonstrated the strongest change between ER– and ER+ breast cancer (fold change = 2.4, p = 1.5E−20). In a correlation analysis with genome-wide expression data in a subcohort of 154 tumors, we found a strong negative correlation (Spearman R = −0.62) between beta-alanine and 4-aminobutyrate aminotransferase (ABAT). Immunohistological analysis confirmed down-regulation of the ABAT protein in ER– breast cancer. In a Kaplan–Meier



Journal of proteome research

Article

Cite This: *J. Proteome Res.* 2018, 17, 337−347

pubs.acs.org/jpr

Deep Learning Accurately Predicts Estrogen Receptor Status in Breast Cancer Metabolomics Data

Fadhl M. Alakwaa,[†] Kumardeep Chaudhary,[†] and Lana X. Garmire[*,†,‡]

[†]Epidemiology Program, University of Hawaii Cancer Center, Honolulu, Hawaii 96813, United States
[‡]Molecular Biosciences and Bioengineering Graduate Program, University of Hawaii at Manoa, Honolulu, Hawaii 96822, United States

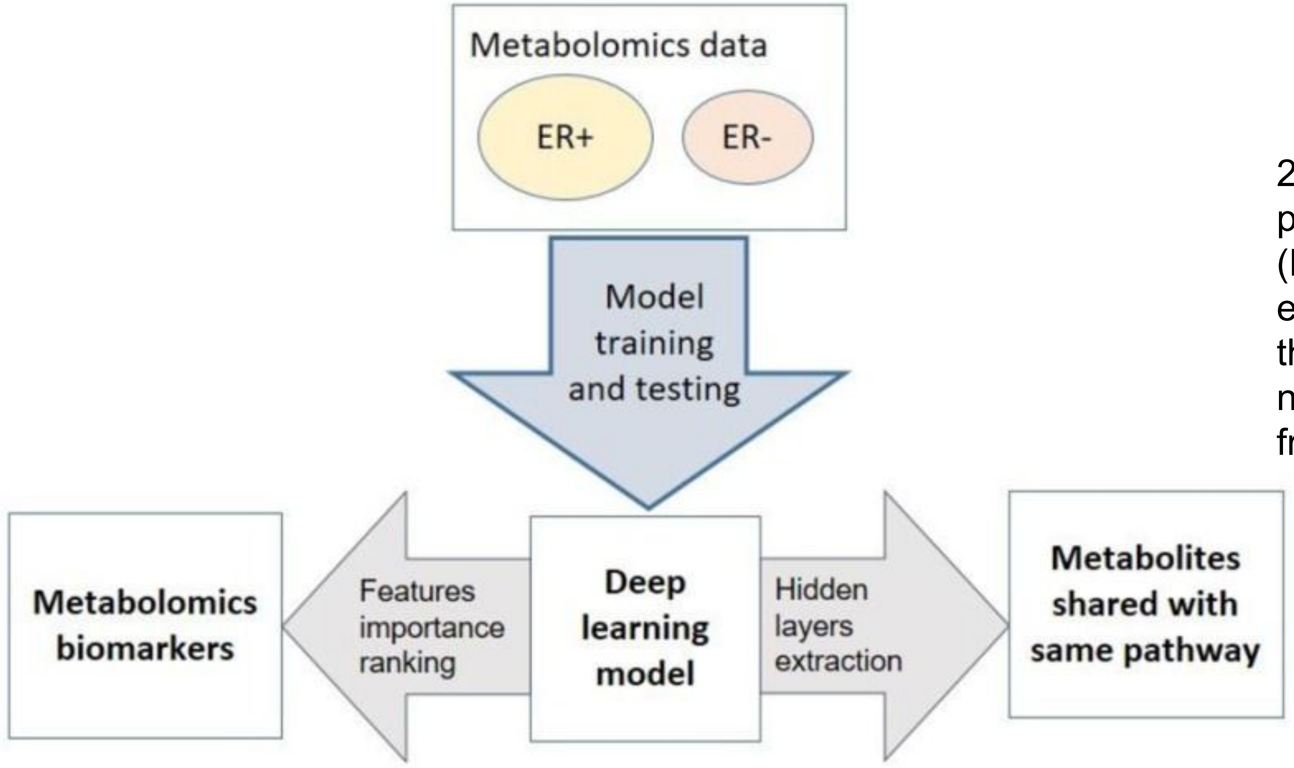
Supporting Information

ABSTRACT: Metabolomics holds the promise as a new technology to diagnose highly heterogeneous diseases. Conventionally, metabolomics data analysis for diagnosis is done using various statistical and machine learning based classification methods. However, it remains unknown if deep neural network, a class of increasingly popular machine learning methods, is suitable to classify metabolomics data. Here we use a cohort of 271 breast cancer tissues, 204 positive estrogen receptor (ER+), and 67 negative estrogen receptor (ER−) to test the accuracies of feed-forward networks, a deep learning (DL) framework, as well as six widely used machine learning models, namely random forest (RF), support vector machines (SVM), recursive partitioning and regression trees (RPART), linear discriminant analysis (LDA), prediction analysis for microarrays (PAM), and generalized boosted models (GBM). DL framework has the highest area under the curve (AUC) of 0.93 in classifying ER+/ER− patients, compared to the other six machine learning algorithms. Furthermore, the biological interpretation of the first hidden layer reveals eight commonly enriched significant metabolomics pathways (adjusted *P*-value <0.05) that cannot be discovered by other machine learning methods. Among them, protein digestion and absorption and ATP-binding cassette (ABC) transporters pathways are also confirmed in integrated analysis between metabolomics and gene expression data in these samples. In summary, deep learning method shows advantages for metabolomics based breast cancer ER status classification, with both the highest prediction accuracy (AUC = 0.93) and better revelation of disease biology. We encourage the adoption of feed-forward networks based deep learning method in the metabolomics research community for classification.

Metabolomics data
ER+
ER-
Model training and testing
Metabolomics biomarkers
Features importance ranking
Deep learning model
Hidden layers extraction
Metabolites shared with same pathway

KEYWORDS: *breast cancer, metabolomics, estrogen receptor, deep learning, bioinformatics*

abstract

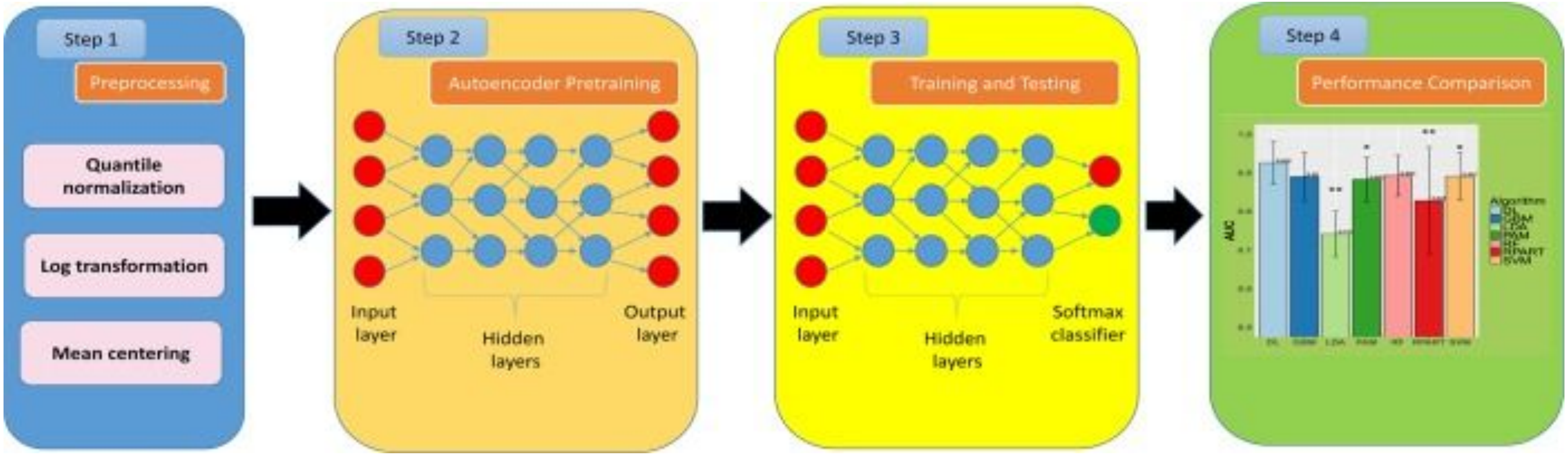

Metabolomics data
ER+
ER-
Model training and testing
Metabolomics biomarkers
Features importance ranking
Deep learning model
Hidden layers extraction
Metabolites shared with same pathway

271 breast cancer tissues, 204 positive estrogen receptor (ER+), and 67 negative estrogen receptor (ER−) to test the accuracies of feed-forward networks, a deep learning (DL) framework

Data Set Information

- The metabolomics data used in this study consists of 271 breast cancer samples (204 ER+ and 67 ER-) collected from a biobank at the Pathology Department of Charité Hospital, Berlin, Germany.
- A total of 162 metabolites with known chemical structure were measured using gas chromatography followed by time-of-flight mass spectroscopy (GC-TOFMS) for all tissue samples.

Work flow



Step 1

Preprocessing

Quantile normalization

Log transformation

Mean centering

Step 2

Autoencoder Pretraining

Input layer

Hidden layers

Output layer

Step 3

Training and Testing

Input layer

Hidden layers

Softmax classifier

Step 4

Performance Comparison

Importing Modules and Dependencies

In any data analysis project, the first step is importing the necessary libraries and modules.

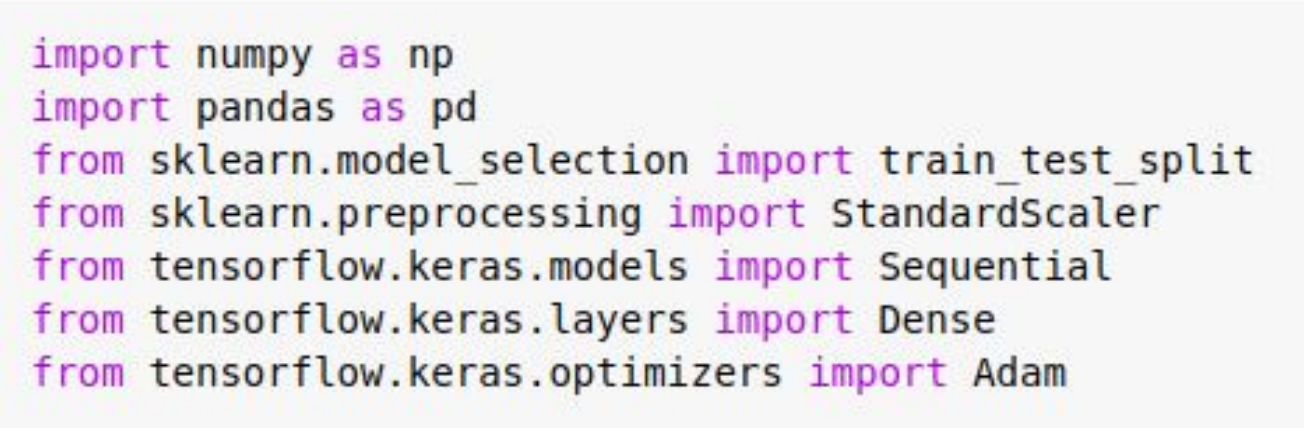
We've used several key libraries in our analysis:

NumPy for numerical operations.

Pandas for data manipulation.

scikit-learn for machine learning tools.

TensorFlow/Keras for deep learning.



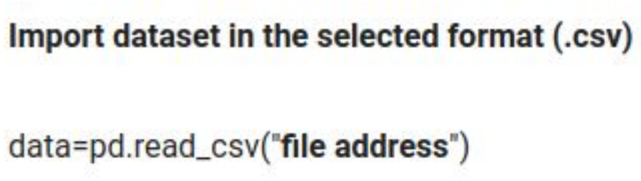
```python
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense
from tensorflow.keras.optimizers import Adam
```

Importing and Loading Data

Data import is a critical step in data analysis.

We imported our dataset from a CSV file located at a specific address.

Data import sets the foundation for subsequent analysis.



Import dataset in the selected format (.csv)

```
data=pd.read_csv("file address")
```

Let’s continue it on Colab sheet!!!

Exploratory Data Analysis (EDA)

EDA helps us understand our data.

We performed the following EDA tasks:

Calculated correlations and visualized them with a heatmap.

Generated histograms to explore data distributions.

Created a word cloud to visualize column names.

Data Preprocessing

Data preprocessing is essential for preparing data for analysis.

Our preprocessing steps included:

Feature extraction.

Normalization using StandardScaler.

Splitting the dataset into training and testing sets.

Autoencoder Neural Network

Autoencoders are neural networks used for feature extraction.

Our autoencoder architecture:

Input layer.

Multiple hidden layers.

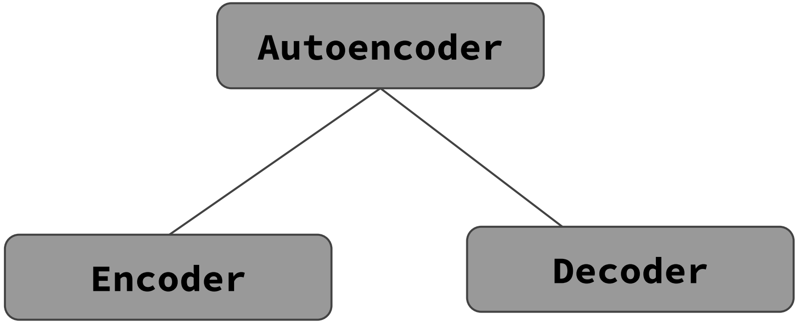
Output layer for feature reconstruction.

We trained the autoencoder to capture essential features in our data.

What Are Autoencoders?

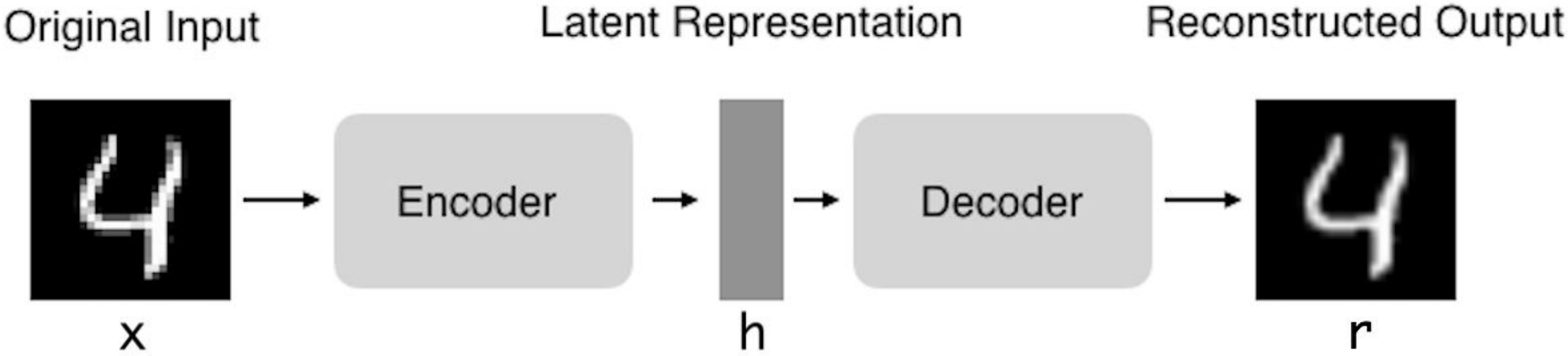
AE is a neural network that is trained to attempt to copy its input to its output.

The network consists of two parts-



$h = f(x)$

$r = g(h)$



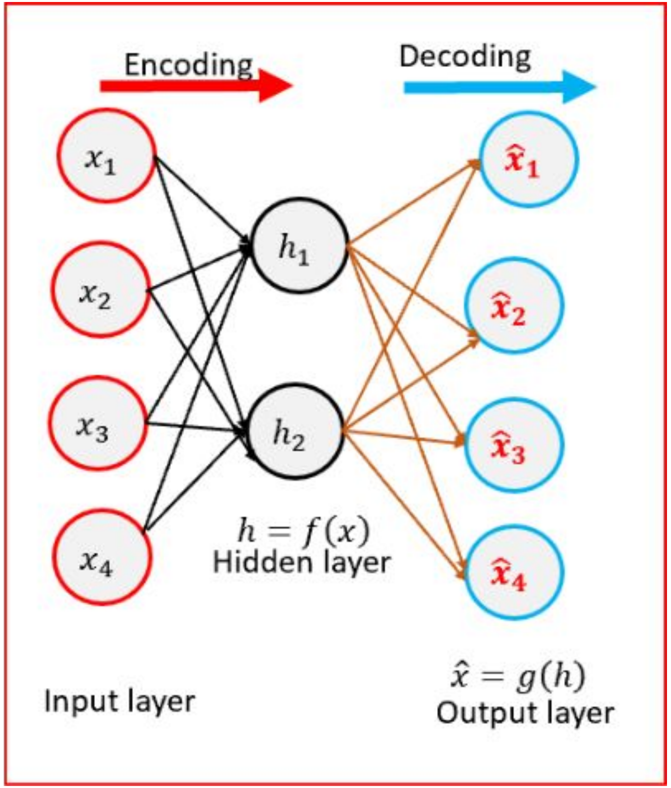
Encoder: compress input into a latent-space of usually smaller dimension. h = f(x)
Decoder: reconstruct input from the latent space. r = g(f(x)) with r as close to x as possible

Autoencoders encodes the input values x, using a function f. It then decodes the encoded values f(x), using a function g, to create output values identical to the input values.

- We take the input, encode it to identify latent feature representation. Decode the latent feature representation to recreate the input.

- We calculate the loss by comparing the input and output. To reduce the reconstruction error.
  - **Autoencoder's objective is to minimize reconstruction error between the input and output. This helps autoencoders to learn important features present in the data.**

- we back propagate and update the weights. Weight is updated based on how much they are responsible for the error.



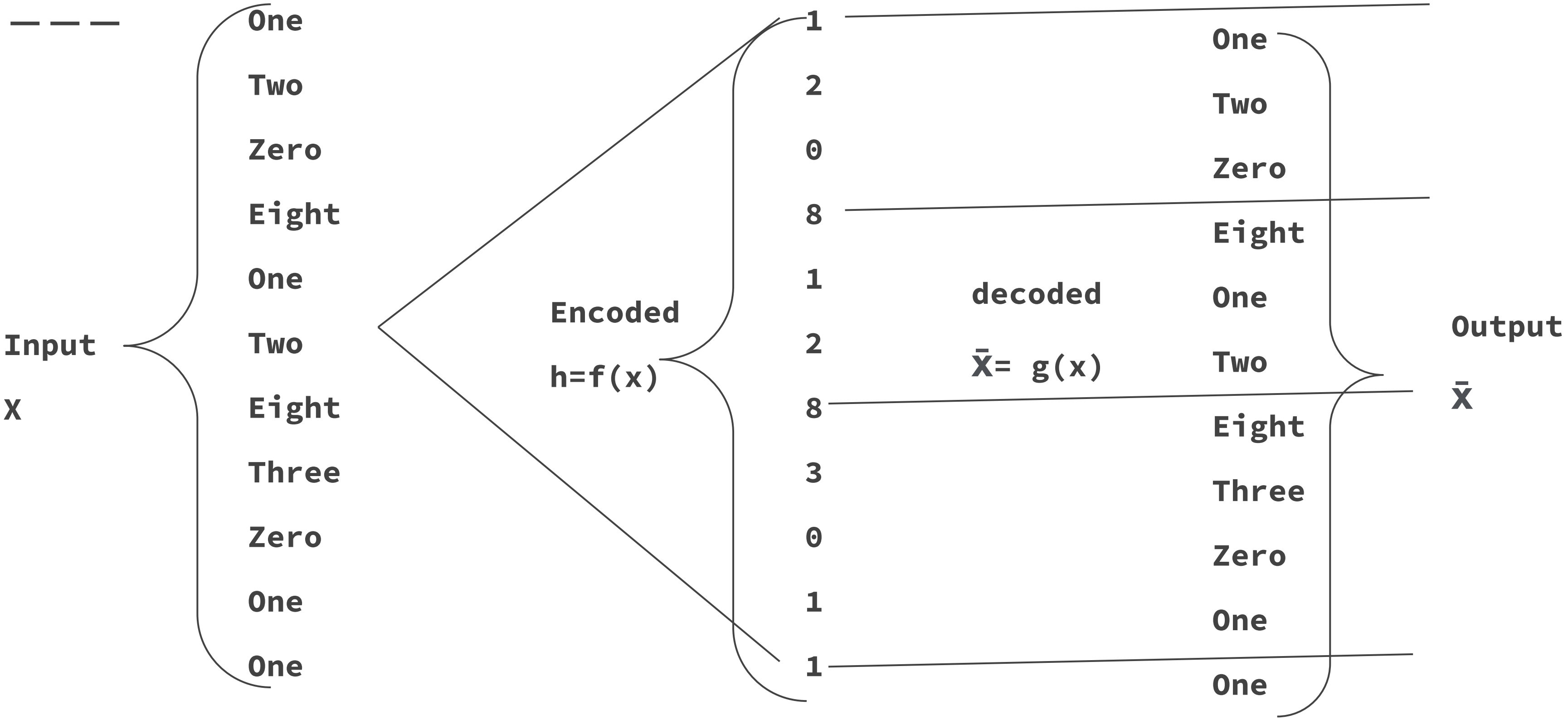
Which one is correct number

```
One
Two
Zero
Eight
One
Two
Eight
Three
Zero
One
Two
```

```
One
Two
Zero
Eight
One
Two
Eight
Three
Zero
One
One
```

```
One
Two
Zero
Eight
One
Two
Eight
Three
Zero
One
three
```

Explanation



Input

X

One

Two

Zero

Eight

One

Two

Eight

Three

Zero

One

One

Encoded

h=f(x)

1

2

0

8

1

2

8

3

0

1

1

decoded

$\bar{X}$= g(x)

One

Two

Zero

Eight

One

Two

Eight

Three

Zero

One

One

Output

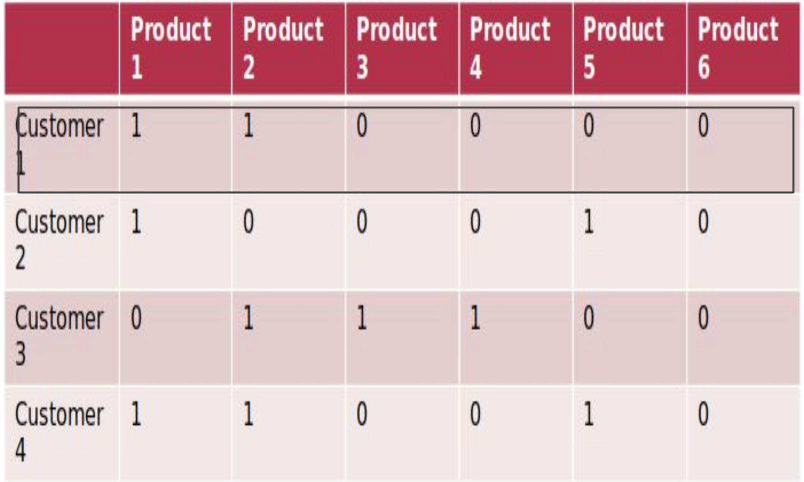
$\bar{X}$

Steps of implementation

Step 1: Take the first row from the customer data for all products bought in an array as the input. 1 represent that the customer bought the product. 0 represents that the customer did not buy the product.

Step 2: Encode the input into another vector *h. h* is a lower dimension vector than the input. We can use sigmoid activation function for h as the it ranges from 0 to 1. W is the weight applied to the input and b is the bias term.

Step 3: Decode the vector *h* to recreate the input. Output will be of same dimension as the input.



	Product 1	Product 2	Product 3	Product 4	Product 5	Product 6
Customer 1	1	1	0	0	0	0
Customer 2	1	0	0	0	1	0
Customer 3	0	1	1	1	0	0
Customer 4	1	1	0	0	1	0

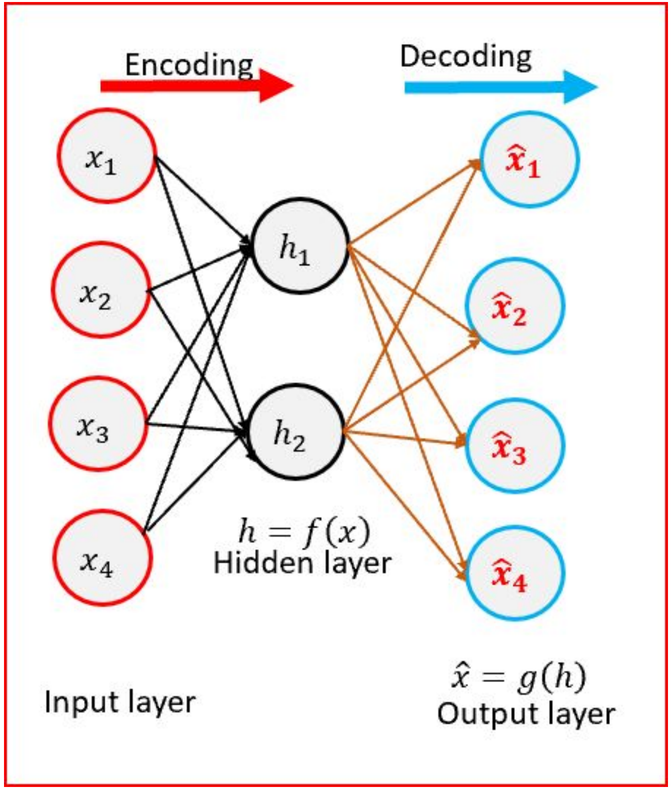
Step 4 : Calculate the reconstruction error L. Reconstruction error is the difference between the input and output vector. Our goal is to minimize the reconstruction error so that output is similar to the input vector.

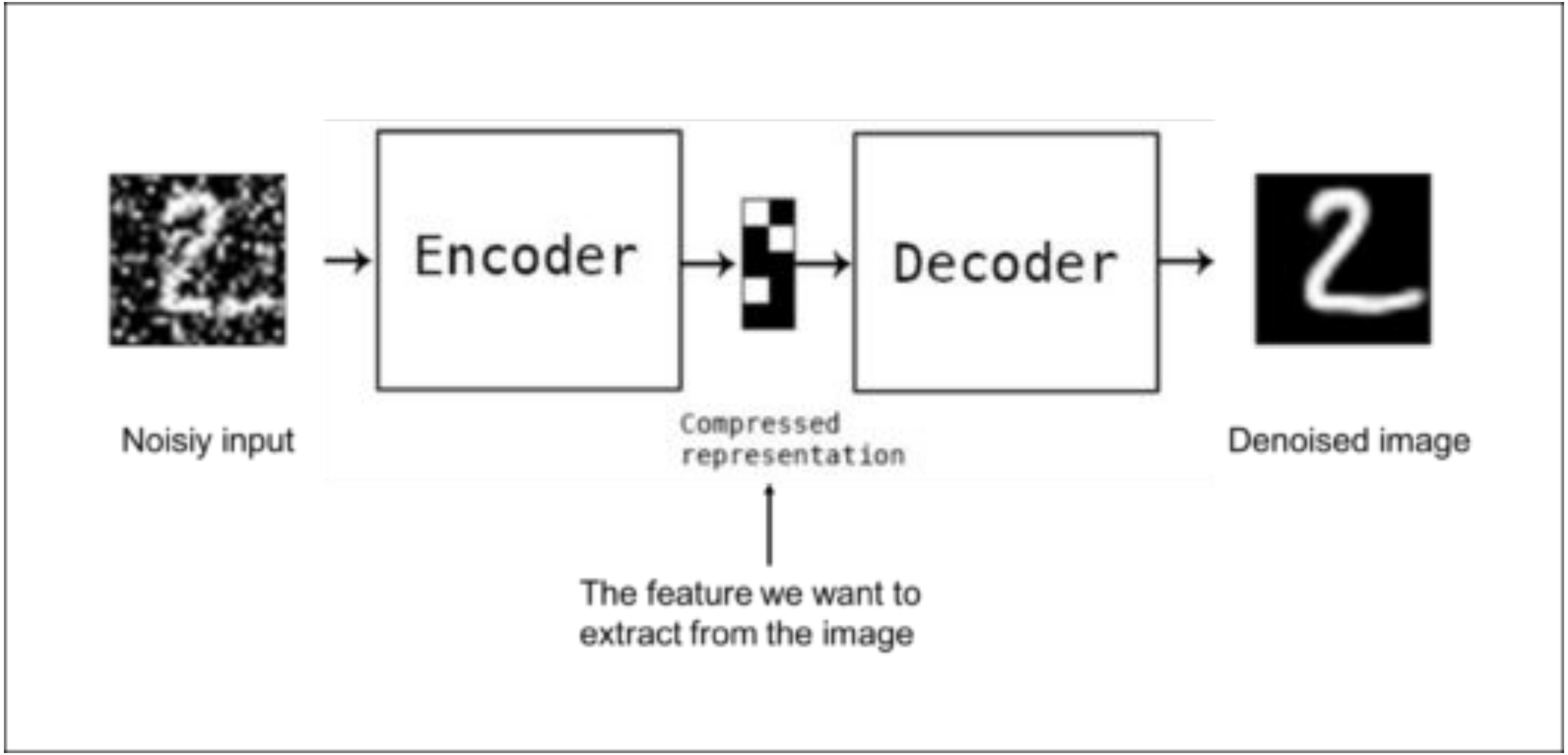
Reconstruction error= input vector — output vector

Step 5: Back propagate the error from output layer to the input layer to update the weights. Weights are updated based on how much they were responsible for the error.

Step 6: Repeat step 1 through 5 for each of the observation in the dataset. Weights are updated after each observation.

Step 7: Repeat more epochs. Epoch is when all the rows in the dataset has passed through the neural network.




Noisiy input
Encoder
Compressed representation
Decoder
Denoised image
The feature we want to extract from the image

Artificial Neural Network (ANN)

Artificial Neural Networks (ANNs) are used for classification.

Our ANN architecture:

Input layer.

Hidden layers for complex pattern recognition.

Output layer for binary classification.

We trained the ANN for classification tasks.

Summary of Results

Let's summarize the results:

Autoencoder extracted essential features.

ANN achieved a certain accuracy.

Conclusion

- In conclusion, metabolomics data analysis with machine learning is a powerful approach.
- These techniques can aid in understanding complex biological systems.
- The choice of model depends on the specific task and data characteristics.